Cognome : **Nome :** **Matricola:**

Modulo: Laboratorio di Metodi Matematici e Statistici, M-Z

Appello del 10 Luglio 2017

Esercizio 1.
Nella seguente tabella sono riportati parzialmente i dati estratti da un campione casuale di 250 dipendenti di una grande multinazionale e relativi al numero di giorni di malattia fatti dal dipendente nel 2016 e al ruolo ricoperto.

	Ruolo			
Giorni	Impiegato	Quadro	Dirigente	
≤ 1	25		8	78
$2-5$	45	32		
>5	30		23	
		89	61	250

a) Completare la tabella inserendo le frequenze corrette negli spazi vuoti. Di che tipo di tabella si tratta?
b) Estraendo un dipendente a caso nel campione, indicare la probabilità che:
b1) il dipendente sia un impiegato;
b2) il dipendente abbia fatto al massimo un giorno di malattia nel 2016;
b3) il dipendente sia un quandro ed abbia fatto più di 5 giorni di malattia nel 2016.
c) Calcolare la probabilità che un dipendente scelto a caso nel campione sia un dirigente sapendo che ha fatto più di 5 giorni di malattia nel 2016.
d) Con riferimento al punto c), i due eventi sono indipendenti? Giustificare la risposta.
e) Sulla base del campione, si può sostenere al livello di significatività dell'1% che il numero di giorni di malattia ed il ruolo ricoperto siano variabili indipendenti?

Soluzione
a) Si tratta di una tabella di contingenza, cioè di una tabella che riporta le frequenze congiunte e marginali rispetto alle due variabili osservate: numero di giorni di malattia nel 2016 (M) e ruolo ricoperto (R). La tabella completa è la seguente:

	Ruolo			
Giorni	Impiegato	Quadro	Dirigente	
≤ 1	25	**45**	8	78
$2-5$	45	32	**30**	**107**
>5	30	**12**	23	**65**
	100	89	61	250

b)
b1) Gli impiegati sono 100 su un totale di 250 e quindi, visto che il dipendente viene scelto *a caso*, la probabilità cercata vale $p_1 = \frac{100}{250} = 0.4 = 40\%$.
b2) I dipendenti che hanno fatto al massimo un giorno di malattia nel 2016 sono 78, quindi $p_2 = \frac{78}{250} = 0.312 = 31.2\%$.
b3) I quadri con più di 5 giorni di malattia sono in totale 12, pertanto $p_3 = \frac{12}{250} = 0.048 = 4.8\%$.
c) I dipendenti con più di 5 giorni di malattia nel 2016 sono 65, di cui 23 dirigenti, quindi la probabilità cercata, una probabilità condizionata, è $p_4 = \frac{23}{65} = 0.354 = 35.4\%$.
d) I due eventi A="il dipendente è un dirigente" e B=" il dipendente ha fatto più di 5 giorni di malattia nel 2016" non sono indipendenti. Infatti: $P(A) = \frac{61}{250} = 0.244$, $P(B) = \frac{65}{250}$ e $P(A \cap B) = \frac{23}{250} = 0.092$, ma $P(A) \times P(B) = \frac{61}{250} \times \frac{65}{250} = 0.06344 \neq 0.092$
e) Per rispondere alla domanda usiamo i dati in tabella per calcolare l'indice del chi-quadrato (χ^2) e

verificare l'ipotesi nulla "le variabili R ed M sono indipendenti" ($H_0: \chi^2 = 0$) contro l'alternativa che non lo siano ($H_1: \chi^2 > 0$). La tabella delle frequenze attese (teoriche) sotto l'ipotesi di indipendenza è data da:

	Ruolo			
Giorni	Impiegato	Quadro	Dirigente	
≤ 1	31.2	27.8	19.0	78
$2-5$	42.8	38.1	26.1	107
>5	26.0	23.1	15.9	65
	100	89	61	250

Si noti che tutti i valori sono superiori a 5. L'indice del chi-quadrato vale $\chi^2 = 29.03$ che risulta superiore a qualunque valore critico di una distribuzione χ^2 a $2 \times 2 = 4$ gradi di libertà (il valore critico al livello $\alpha = 0.0001$ vale 23.5064, inferiore a 29.03). E quindi si rifiuta l'ipotesi di indipendenza a qualunque grado di significatività.

Esercizio 2
Per uno studio su una patologia cardio–vascolare viene misurato l'indice di massa corporea (in kg/m^2) in un campione casuale di 15 individui, ottenendo i seguenti valori:

$$31.6,\ 34.9,\ 43.1,\ 28.6,\ 49.2,\ 31.1,\ 27.5,\ 24.5,\ 19.6,\ 34.6,\ 42.7,\ 41.5,\ 38.6,\ 21.2,\ 28.0$$

a) Determinare l'intervallo di confidenza al livello del 95% per il valore medio dell'indice di massa corporea nella popolazione di riferimento. Quali ipotesi sono necessarie per calcolare tale intervallo?
b) Con riferimento al punto a), quanto vale il margine d'errore? Senza fare calcoli, spiegare come varia il margine d'errore rispetto al livello dell'intervallo.
c) Secondo l'Organizzazione Mondiale della Sanità, un indice di massa corporea superiore a 30 kg/m^2 è indicatore di obesità. Sulla base del campione, c'è abbastanza evidenza per affermare che la popolazione di riferimento è in media obesa?
d) A parità di media e varianza campionarie, quanto dovrebbe essere grande il campione per avere evidenza a favore dell'obesità media della popolazione al livello di significatività dell'1%?

Soluzione
a) Si ha $\bar{x} = 33.113$ e $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 74.240$. Sotto l'ipotesi che la variabile casuale X ="indice di massa corporea" abbia una distribuzione gaussiana con media μ e varianza σ^2 incognite, l'intervallo cercato è: $\text{IC}(0.95) = (\bar{x} - t(14)_{0.025}\sqrt{s^2/15}, \bar{x} + t(14)_{0.025}\sqrt{s^2/15}) = (33.113 - 2.1448\sqrt{74.24/15}, 33.113 - 2.1448\sqrt{74.24/15}) = (28.341, 37.884)$
b) il margine d'errore è dato dalla semi–ampiezza dell'intervallo: $t(14)_\alpha\sqrt{s^2/15} = 4.771$. Per $s^2 = 74.24$, all'aumentare del livello $1-\alpha$ dell'intervallo aumenta anche $t(14)_\alpha$ e, pertanto, aumenta anche il margine di errore.
c) Sottoponiamo a verifica l'ipotesi nulla $H_0: \mu = 30$ contro l'alternativa che in media la popolazione di riferimento sia obesa, $H_1: \mu > 30$. La statistica test vale

$$\frac{\bar{x} - 30}{\sqrt{s^2/n}} = \frac{33.113 - 30}{\sqrt{74.24/15}} = 1.399 < 1.7613 = t(14)_{0.05}$$

e pertanto non c'è abbastanza evidenza nel campione per sostenere che la popolazione di riferimento sia mediamente obesa al livello del 5%.
d) Si tratta di stabilire il valore di n per cui

$$\frac{\bar{x} - 30}{\sqrt{s^2/n}} = \frac{33.113 - 30}{\sqrt{74.24/n}} = \sqrt{n}\,\frac{33.113 - 30}{\sqrt{74.24}} > t(n)_{0.01}\ .$$

Ricordando che per $n > 30$ la distribuzione $t-Student$ è approssimabile con la distribuzione gaussiana, possiamo considerare la disequazione:

$$\sqrt{n}\,\frac{33.113 - 30}{\sqrt{74.24}} > z_{0.01}$$

e risolvere rispetto a n:

$$\sqrt{n} > \frac{\sqrt{74.24}}{33.113 - 30} \times 2.3263 = 6.439$$

da cui, elevando al quadrato, si ottiene $n > 41.5$. Quindi, a parità di media e varianza campionarie, se la dimensione del campione fosse stata di almeno 42 unità ci sarebbe stata abbastanza evidenza nel campione a favore dell'obesità.

Esercizio 3

Per uno studio su una patologia cardio–vascolare viene scelto un campione casuale di 10 individui a cui vengono rilevati l'indice di massa corporea, BMI, (in kg/m^2) e un indice, AHI, che descrive la gravità di una malattia, la sindrome delle apnee ostruttive nel sonno (OSA). I risultati ottenuti sono riportati nella seguente tabella:

BMI	21.1	35.3	21.4	33.0	27.8	39.7	45.3	26.0	44.3	30.3
AHI	7.3	68.7	1.3	40.1	22.0	12.8	67.4	31.2	72.9	24.9

a) Lo studio si propone di dimostrare che il peso è un fattore di rischio per la sindrome OSA, cioè che all'aumentare del peso cresce anche la gravità della sindrome. Quale dei due grafici sottostanti è il più adatto per condurre lo studio? Giustificare la risposta.

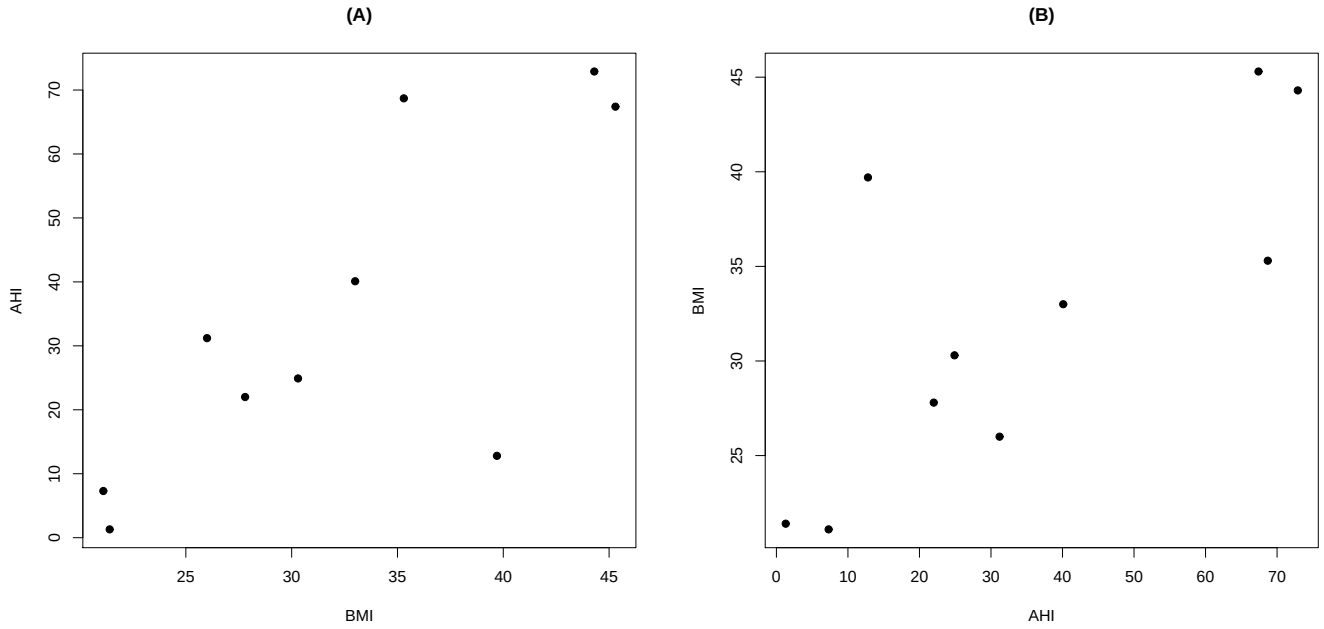

Figura 1: Esercizio 3

b) Con riferimento alla scelta fatta nel punto a), fornire una stima dei parametri di un modello di regressione lineare semplice tra BMI e AHI e rappresentare il modello stimato sul grafico scelto.
c) Testare la significatività della regressione al livello del 5%.
d) Per quali dei seguenti valori di BMI la previsione risulta più attendibile: BMI=15, BMI=30, BMI=43? Giustificare la risposta.

Soluzione.

a) Il grafico più adatto è il grafico (A), in cui la variabile indipendente è il peso (BMI) mentre la variabile dipendente (la *risposta*) è la gravità della sindrome (AHI). Per comodità indichiamo con X la variabile BMI e con Y la variabile AHI.
b) Per calcolare le stime dei parametri della retta $y = a + bx$ servono: $\bar{x} = 32.42$, $\bar{y} = 34.86$, $\sigma_x^2 = (n^{-1}\sum_i x_i^2) - \bar{x}^2 = 68.53$, e $\sigma_{xy} = (n^{-1}\sum_i x_i y_i) - \bar{x}\,\bar{y} = 159.677$ da cui:

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{159.677}{68.53} = 2.33 \quad \hat{a} = \bar{y} - \hat{b}\bar{x} = 34.86 - 2.33 \times 32.42 = -40.679.$$

Ricordando che la retta di regressione passa per il baricentro dei dati $(\bar{x},\ \bar{y}) = (32.42, 34.86)$ e, per esempio, per il punto di coordinate $x = 25$ e $y = -40.679 + 2.33 \times 25 = 17.571$, la retta di regressione stimata è quella in Figura 2.

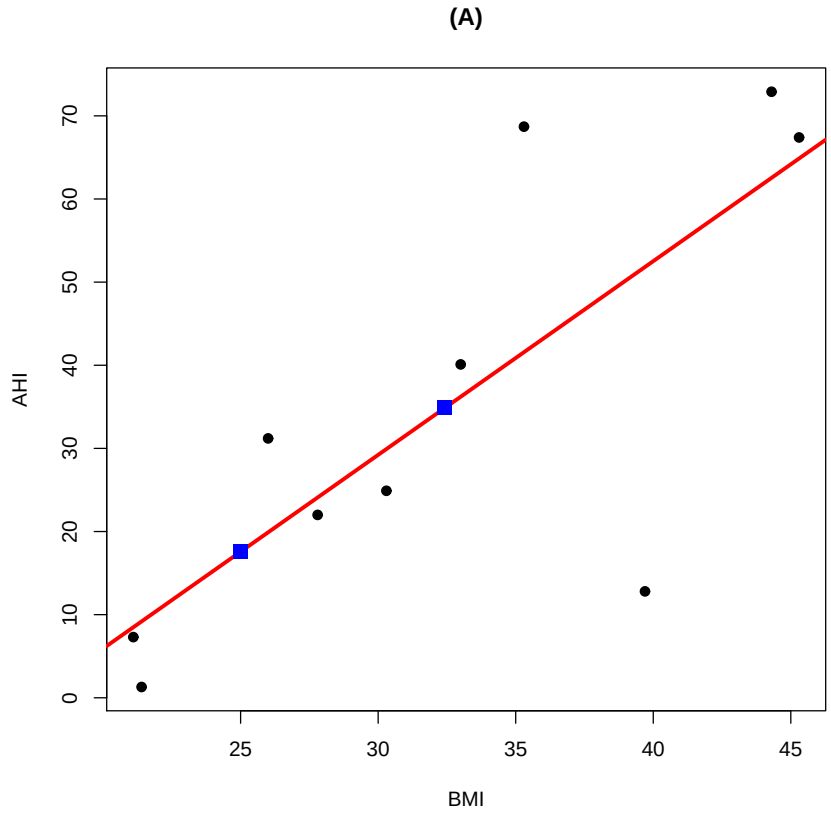

Figura 2: Esercizio 3

c) Sottoponiamo a verifica l'ipotesi nulla $H_0:\ b = 0$ contro l'alternativa $H_1:\ b \neq 0$. La statistica test è

$$\frac{|\hat{b}|}{\sqrt{\frac{s^2}{n\sigma_x^2}}}$$

e quindi ci serve calcolare s^2, la la stima della varianza dell'errore $\epsilon_i \sim N(0, \sigma^2)$ del modello lineare $y_i = a + bx_i + \epsilon_i$. Per questo ci basta calcolare $\sigma_y^2 = (n^{-1} \sum_i y_i^2) - \bar{y}^2 = 632.894$. Infatti:

$$R^2 = \rho^2 = \left(\frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} \right)^2 = 0.588$$

(che, per inciso, indica una bontà di adattamento del modello ai dati sotto la soglia indicativa di 0.7) e

$$s^2 = \frac{n}{n-2} \times (1 - R^2) \times \sigma_y^2 = \frac{10}{8} \times (1 - 0.588) \times 632.894 = 325.94$$

La statistica test vale, quindi, $2.33/\sqrt{325.94/(10 \times 68.53)} = 3.377 > t(8)_{0.025} = 2.306$ e, pertanto, la regressione è significativa al livello del 5%.
d) Si ricordi che la retta di regressione ha validità solo nell'intervallo definito dalla variabile indipendente, qui BMI. Siccome BMI varia da 21.1 a 45.3, non possiamo usare la retta stimata per prevedere AHI quando BMI=15. La previsione è tanto più attendibile quanto più il valore è vicino a $\bar{x}$=32.42, perchè l'intervallo di confidenza per la previsione nel punto x_0 dipende direttamente da $(x_0 - \bar{x})^2$ e quindi la previsione sarà più attendibile per BMI=30 che per BMI=43.